

The Rich Transcription 2005 Spring Meeting Recognition Evaluation

Jonathan G. Fiscus¹, Nicolas Radde¹, John S. Garofolo¹, Audrey Le¹, Jerome Ajot¹,
Christophe Laprun^{1,2}

¹National Institute Of Standards and Technology, 100 Bureau Drive Stop 8940, Gaithersburg, MD 20899

² Systems Plus, Inc., 1370 Piccard Drive, Suite 270, Rockville, MD 20850
{jfiscus,nradde,jgarofolo,ajot,claprun}@nist.gov

Abstract. This paper presents the design and results of the Rich Transcription Spring 2005 (RT-05S) Meeting Recognition Evaluation. This evaluation is the third in a series of community-wide evaluations of language technologies in the meeting domain. For 2005, four evaluation tasks were supported. These included a speech-to-text (STT) transcription task and three diarization tasks: “Who Spoke When”, “Speech Activity Detection”, and “Source Localization.” The latter two were first-time experimental proof-of-concept tasks and were treated as “dry runs”. For the STT task, the lowest word error rate for the multiple distant microphone condition was 30.0% which represented an impressive 33% relative reduction from the best result obtained in the last such evaluation - the Rich Transcription Spring 2004 Meeting Recognition Evaluation. For the diarization “Who Spoke When” task, the lowest diarization error rate was 18.56% which represented a 19% relative reduction from that of RT-04S.

1. Motivation

The National Institute of Standards and Technology (NIST) has been working with the speech recognition community since the mid 1980s to improve the state-of-the-art in technologies for transforming speech into text. To facilitate progress, NIST has worked with the community to make training/development data collections available for several speech domains. NIST collaborated with the research community to define performance metrics and create evaluation tools so that technology developers can perform hill-climbing experiments and self-evaluate their progress. NIST also coordinates periodic community-wide benchmark tests and technology workshops to inform the research community and Government sponsors of progress and to promote technical exchange. The test suites used in these benchmark tests are generally made available to the community as development tools after the formal evaluations.

NIST’s evaluations have demonstrated great progress in the state-of-the-art in speech-to-text (STT) transcription systems. STT systems in the late 80s focused on read speech from artificially-constrained domains. As the technology improved, the NIST evaluations focused the research community on increasingly difficult challenges with regard to speech modality, speaker population, recording characteristics, language, vocabulary, etc. Now that English Broadcast News word error rates are below

10% and English Conversational Telephone Speech word error rates are nearing 15% [1], it is apparent that the research community is ready for the next challenge.

The meeting domain presents several challenges to the technology which aren't represented in the broadcast news and conversational telephone speech domains. These include varied forums and an infinite number of topics, spontaneous highly interactive and overlapping speech, varied recording environments, varied/multiple microphones, multi-modal inputs, participant movement, and far field speech effects like ambient noise and reverberation. In order to properly study these challenges, laboratory-quality experiment controls must be available to enable systematic research. The meeting domain provides a unique environment to collect naturally-occurring spoken interactions under controlled sensor conditions.

The Rich Transcription Spring 2005 (RT-05S) Meeting Recognition evaluation is part of the NIST Rich Transcription (RT) series of language technology evaluations [1] [2]. These evaluations have moved the technology focus from a strictly word-centric approach to an integrated approach where the focus is on creating richly annotated transcriptions of speech, of which words are only one component. The goal of the RT series is to create technologies to generate transcriptions of speech which are fluent and informative and which are readable by humans and usable in downstream processing by machines. To accomplish this, lexical symbols must be augmented with important informative non-orthographic metadata. These resulting metadata enriched transcripts are referred to as "rich transcriptions." This approach was originated in the DARPA Effective, Affordable, Reusable Speech-to-Text (EARS) Program¹ and is being continued by NIST and other research communities. These metadata can take many forms (e.g., which speakers spoke which words, topic changes, syntactic boundaries, named entities, speaker location, etc.)

The RT-05S evaluation is the result of a multi-site/multi-national collaboration. In addition to NIST, the organizers and contributors included: the Augmented Multiparty Interaction (AMI) program, the Computers in the Human Interaction Loop (CHIL) program, Carnegie Mellon University (CMU), Evaluations and Language resources Distribution Agency (ELDA), International Computer Science Institute and SRI International (ICSI/SRI), The Center for Scientific and Technological Research (ITC-irst), Karlsruhe University (KU), the Linguistic Data Consortium (LDC), and Virginia Tech (VT). AMI, CMU [9], ICSI [7], NIST [8], and VT each donated two meetings recorded at their labs to the evaluation. Excerpts from these meetings were selected to comprise the RT-05 conference room test set which is similar in design to the RT-04S test set. KU donated sixteen meetings to make a separate lecture room test set. CMU, ITC-irst, KU, LDC, and ELDA collaborated to prepare the reference transcripts and annotations.

¹ <http://www.darpa.mil/ipto/Programs/ears/index.htm>

2. Rich Transcription Spring 2005 Meeting Recognition Evaluation

The RT-05S evaluation broke ground on four fronts. First, new audio sensors and digital microphone arrays were added to the test conditions. Second, a new STT evaluation tool developed at NIST was released to the participants to score transcriptions of simultaneous overlapping speech. Third, two test sets were prepared for the evaluation, each representing two meeting sub-domains: small conference room meetings and lectures. Fourth, the conference room test set contained two meetings from Virginia Tech for which no training data was available.

All participating teams were required to submit a single primary system on the required task-specific evaluation condition. The primary systems are expected, by the developers, to be their best performing systems. NIST's analysis focuses on these primary systems.

The Rich Transcription Spring 2005 Evaluation plan [3] documents the Rich Transcription Spring 2005 (RT-05S) Meeting Recognition evaluation. The evaluation plan describes in detail the evaluation tasks, data sources, microphone conditions, system input and output formats, and evaluation metrics employed in the evaluation. This section summarizes the evaluation plan and covers the meeting sub-domains represented in the test set, the audio input conditions supported by the test corpora, and the evaluation task definitions

2.1 Meeting Sub-Domains: Conference Room vs. Lecture Room

The meeting domain is highly variable along several dimensions. In the broad sense, any interaction between 2 more people may be considered a meeting. As such, meetings can range from brief informal exchanges to extremely formal proceedings with many participants following specific rules of order. There are a number of factors that shape how the participants interact with each other. Further, it is well known that the type, number, and placement of sensors have a significant impact on the performance of recognition tasks. The variability is so large that it would be impossible to build either a training or testing corpus that encompasses all of these factors. To make the problem tractable, the RT evaluations have attempted to constrain the definition to two specific sub-domains: small conference room meetings (also occasionally referred to as "board room" meetings) and "lecture room" meetings. The two sub-domains are used to differentiate between two very different participant interaction modes as well as two different sensor setups. The RT-05S evaluation includes a separate test set for each of these two sub-domains, labeled "confmtg" and "lectmtg."

In addition to differences in room and sensor configuration, the primary difference between the two sub-domains is in the group dynamics of the meetings. The RT conference meetings are primarily goal-oriented decision-making exercises and are either moderated or lead by one member of the meeting. As such, these meetings are highly-interactive and multiple participants contribute to the information flow and decisions made. In contrast, lecture meetings are educational events where a single lecturer is briefing the audience on a particular topic. While the audience occasionally

participates in question and answer periods, it rarely controls the direction of the interchange or the outcome.

Section 2.4 describes the corpora used for both the *lectmtg* and *confmtg* domains in the RT-05S evaluation.

2.2 Microphone Conditions

As with RT-04S, three core input conditions were supported for RT-05S: multiple distant microphones (MDM), single distant microphone (SDM), and individual head microphones (IHM). The troika of audio input conditions makes a very powerful set of experimental controls for black box evaluations. The MDM condition provides a venue for the demonstration of multi-microphone input processing techniques. It lends itself to experimenting with simple beamforming and noise abatement techniques to address room acoustic issues. The SDM input condition provides a control condition for testing the effectiveness of multi-microphone techniques. The IHM condition provides two important contrasts: first, it effectively eliminates the effects of room acoustics, background noise, and overlapping simultaneous speech, and second it is most similar to the Conversational Telephone Speech (CTS) domain [1] and may be compared to results in comparable CTS evaluations.

The enumeration below contains definitions of the three previously mentioned audio input conditions and three new microphone sources for the RT-05S evaluation: multiple Mark III microphone arrays, multiple beamformed signals, and multiple source localization arrays.

- Multiple distant microphones: (MDM) This evaluation condition includes the audio from at least 3 omni-directional microphones placed (generally on a table) between the meeting participants. This condition was supported in both the *confmtg* and *lectmtg* datasets.
- Single distant microphone: (SDM) This evaluation condition includes the audio of a single, centrally located omni-directional microphone for each meeting. This microphone channel is selected from the microphones used for the MDM condition. Based on metadata provided with the recordings, it is selected so as to be the most centrally-located omni-directional microphone. This condition was supported in both the *confmtg* and *lectmtg* datasets.
- Individual head microphone: (IHM) This evaluation condition includes the audio recordings collected from a head mounted microphone positioned very closely to each participant's mouth. The microphones are typically cardioid or super cardioid microphones² and therefore the best quality signal for each speaker. Since the IHM condition is a contrastive condition, systems can also use any of the microphones used for the MDM condition. This condition was supported in both the *confmtg* and *lectmtg* datasets.
- Multiple Mark III microphone arrays: (MM3A) This evaluation condition includes audio from all the collected Mark III microphone arrays. The *lectmtg* dataset con-

² After the evaluation began, NIST discovered some of their head microphones were omni-directional.

tains the data from each channel of one Mark-III microphone array per meeting. In addition, the NIST subset of the *confmtg* data contains the data from each channel of three Mark-III microphone arrays per meeting. Due to time constraints, no results were submitted using these data.

- Multiple Source Localization microphone arrays (MSLA): This evaluation condition includes the audio from all the CHIL source localization arrays (SLA). An SLA is a 4-element digital microphone array arranged in an upside down ‘T’ topology [4]. The lecture room meeting recordings include four SLAs, one mounted on each wall of the room.

2.3 Evaluation tasks

Four evaluation tasks were supported for the RT-05S evaluation: a Speech-To-Text transcription task and three diarization tasks: “Who Spoke When”, “Speech Activity Detection”, and “Source Localization.” The latter two tasks were proposed for inclusion by the CHIL program and they were considered dry run tasks for the RT-05S evaluation. The following is a brief description of each of the evaluation tasks:

Speech-To-Text (STT) Transcription: STT systems are required to output a transcript of the words spoken by the meeting participants along with the start and end times for each recognized word. For this task, no speaker designation is required. Therefore, the speech from all participants is to be transcribed as a single word output stream.

Systems were evaluated using the Word Error Rate (WER) metric. WER is defined to be the sum of system transcription errors, (word substitutions, deletions, and insertions) divided by the number of reference words and expressed as a percentage. It is an error metric, so lowers scores indicate better performance. The score for perfect performance is zero. Since insertion errors are counted, it is possible for WER scores to exceed one hundred percent.

WER is calculated by first harmonizing the system and reference transcript through a series of normalization steps. Then the system and reference words are aligned using a Dynamic Programming solution. Once the alignment mapping between the system and reference words is determined, the mapped words are compared to classify them as either correct matches, inserted system words, deleted reference words, or substituted system words. The errors are counted and statistics are generated.

The MDM audio input condition was the primary evaluation condition for the STT task for both meeting sub domains. The *confmtg* data supported two contrastive conditions, SDM and IHM, and the *lectmtg* data supported four contrastive conditions, SDM, IHM, MSLA, and MM3A. Participants could submit systems for the *confmtg* domain, the *lectmtg* domain, or both the sub domains. Systems could use the knowledge of the domain as side information and therefore configure their systems for each sub domain.³

³ All systems, for all tasks, were privy to side information about the data being processed. The evaluation plan enumerates these in detail.

Diarization “Who Spoke When” (SPKR) SPKR systems are required to annotate a meeting with regions of time indicating when each meeting participant is speaking and clustering the regions by speaker. It is a clustering task as opposed to an identification task since the system is not required to output a name for the speakers – only a generic id.⁴

The Diarization Error Rate (DER) metric is used to assess SPKR system performance. DER is the ratio of incorrectly attributed speech time, (either falsely detected speech, missed detections of speech, or incorrectly clustered speech) to the total amount of speech time, expressed as a percentage. As with WER, a score of zero indicates perfect performance and higher scores indicate poorer performance than lower scores.

DER is calculated by first computing a 1:1 mapping between the system-generated speaker clusters and the segment clusters in the reference transcript using the Hungarian solution to a bipartite graph⁵. Once the mapping is found, system segments not attributed to the mapped reference speaker cluster are declared incorrectly clustered speech. Falsely detected speech and missed detections of speech are calculated by simple accumulating the amount of time for each class of error.

For 2005, the primary measure of DER was calculated for non-overlapping speech only in order to be comparable with previous evaluations of speaker diarization. However, given the shifting focus to evaluation of all speech (including overlapping speech), the DER was also computed for overlapping speech segments. Both sets of scores are provided. In future such evaluations, the primary measure will focus on all speech.

Inherent ambiguities in pinpointing speech boundaries in time and annotator variability result in a small degree of inconsistency in the time annotations in the reference transcript. As such, a 0.25 second collar around each reference segment is not scored. This collar effectively minimizes the amount of DER error due to reference annotation inconsistencies.

Another challenge is in determining how large a pause in speech must be to cause a segment break. Although somewhat arbitrary, the cutoff value of 0.3 seconds was empirically determined to be a good approximation of the minimum duration for a pause in speech resulting in an utterance boundary. As such, segments that are closer than 0.3 seconds apart are merged in both the reference and system output transcripts.

The MDM audio input condition was the primary evaluation condition for the SPKR task for both meeting sub domains. The *confmtg* data supported one contrastive condition, SDM, and the *lectmtg* data supported three contrastive conditions, SDM, MSLA, and MM3A. Participants could submit systems for the *confmtg* domain, the *lectmtg* domain, or both the sub domains. Systems could use the knowledge

⁴ In a real meeting transcription application, it is likely that the SPKR and STT system outputs would be merged to attribute each transcribed word to a particular meeting participant. The decision was made to not yet evaluate an integrated STT/SPKR task since, at this early stage, it is important to understand how each of the core components of such a system behaves. It's anticipated that such an integrated task will be included in the RT evaluations in the future.

⁵ <http://www.nist.gov/dads/HTML/HungarianAlgorithm.html>

of the domain as side information and therefore configure their systems for each sub domain.⁶

Diarization “Speech Activity Detection” (SAD) SAD systems are required to annotate a meeting with regions of time indicating when at least one person is talking. The SAD task is therefore a simplified version of the SPKR task because no speaker clustering is performed by the system. The task was introduced to lower barriers for participation and to gauge the contribution of SAD errors to the SPKR and STT tasks. Since this is the first time the SAD task has been included in the RT evaluations, it was treated as an experimental dry run.

Because SAD is viewed as a simplification of the SPKR task, the SPKR DER scoring metric is also used to score the SAD task. The same no-score collar, 0.25 seconds, was applied during scoring and the same smoothing parameter, 0.3 seconds, was applied to the reference files. The reference files were derived from the SPKR reference files by simply merging the reference speaker clusters into a single cluster and then merging segments that either overlap or were within the 0.3 second smoothing parameter.

The MDM audio input condition was the primary evaluation condition for the SAD task for both meeting sub domains. The *confmtg* data supported two contrastive conditions, SDM and IHM, and the *lectmtg* data supported four contrastive conditions, SDM, IHM, MSLA, and MM3A. Participants could submit systems for the *confmtg* domain, the *lectmtg* domain, or both the sub domains. Systems could use the knowledge of the domain as side information and therefore configure their systems for each sub domain.⁷

The SAD task using IHM data is not directly comparable to SAD on distant microphone data, (i.e., MDM, SDM, MSLA, or MM3A data). An IHM channel includes both the wearer’s speech and cross talk for other meeting participants. This cross talk is not considered detectable speech even though it was human generated. Not only must IHM SAD systems detect speech, but also detect when the speech is cross talk. This of course is a much harder problem.

One issue arose during this evaluation regarding meeting participants who speak very little. Since the DER was measured separately for each close-talking microphone and since the denominator of the DER metric is the amount of speech uttered, the DER for quiet speakers may be dominated by falsely detected speech errors. Time did not permit us to examine alternative scoring techniques which would minimize this effect prior to the MLMI workshop.

Diarization “Source Localization” (SLOC): SLOC systems are required to emit the three-dimensional position (in millimeters) of each person who is talking. The labels do not include the speaker’s identity, but systems must be able to distinguish between time periods with speech and without speech. As such, this task is similar to the SAD task with the additional requirement of speaker location. The RT-05S instantiation of

⁶ All systems, for all tasks, were privy to side information about the data being processed. The evaluation plan enumerates these in detail.

⁷ All systems, for all tasks, were privy to side information about the data being processed. The evaluation plan enumerates these in detail.

the SLOC task is a simplified proof-of-concept version of this task. For RT-05S, the task was constrained to segments which contained only a single speaker -- a lecturer in the CHIL lecture room data. Therefore, SLOC for overlapping speakers was not addressed.

In order to infer location from the audio stream, SLOC systems use the source localization array audio data. The three dimensional position of the each SLA microphone element has been computed and given to the systems and from that information; the SLOC systems infer the location of the speaker.

The primary, and only, evaluation condition for the SLOC task is the MSLA audio input condition for the *lectmtg* data. No other audio input conditions were supported for this task and none of the *confmtg* data has SLA recordings.

The definition of the task and evaluation metrics are documented in the CHIL “Speaker Localization and Tracking – Evaluation Criteria” document [4]. The metric used to evaluate the SLOC task was the Root-Mean-Squared of Localization Error (RMSE). The EVAL_IRST_SP_LOC scoring software was developed by ITC-irst and contributed to NIST in scoring the results of the evaluation. The RMSE metric determines the Euclidean distance between the reference speaker position and the system-hypothesized speaker position every 10 milliseconds. The task was added to the evaluation at the request of the CHIL program participants and was supported in large part by the CHIL Program.

2.4 RT-05S Evaluation Corpora Details

As indicated previously, the RT-05S evaluation data consisted of two test sets: a conference room meeting (*confmtg*) test set and a lecture room meeting (*lectmtg*) test set. The recordings were sent to participants as either down sampled 16-bit, 16Khz NIST Speech Header Resources (SPHERE) files or in the original sample format 24-bit, 44.1 Khz WAV and headerless raw files. The recordings of the meetings in the *confmtg* data set were distributed in their entirety while only the selected excerpts from the *lectmtg* data were distributed.⁸ Some of the meeting recordings also included video recordings. However, they were not distributed for the evaluation since none of the evaluation participants planned to implement multi-modal experiments this year. The video recordings may be made available at a later date for future multi-modal system development.

Conference Room Meetings: The *confmtg* test set consisted of nominally 120 minutes of meeting excerpts from ten different meetings. Five sites each provided two meetings for the evaluation test set and NIST selected a twelve minute excerpt from each meeting to be evaluated. The five contributing sites were the Augmented Multi-party Interaction (AMI) Project, Carnegie Mellon University (CMU), the International Computer Science Institute (ICSI), the National Institute of Standards and Technology (NIST), and Virginia Tech (VT). The Linguistic Data Consortium (LDC) transcribed the test set according to the “Meeting Data Careful Transcription Specifica-

⁸ The 1.2 terabytes of lecture meeting data proved too large to distribute within the evaluation time constraints.

tion - V1.2” guidelines [5]. Table 1 gives the salient details concerning the *confmtg* evaluation corpus.

Each meeting recording evaluation excerpt met minimum sensor requirements. Each meeting participant wore a head-mounted close talking microphone and there were at least three table top microphones placed on a table between the meeting participants. The only exception to this is meeting NIST_20050412-1303 from NIST in which a meeting participant was talking over a conference phone. In addition to these sensors, the AMI meetings included an eight-channel circular microphone array placed on the table between the meeting participants, and the NIST meetings included three Mark III arrays mounted on the walls.

Table 1 Summary of Conference Room Meeting evaluation corpus

Meeting ID	Duration (minutes)	Number of Participants	Notes
AMI_20041210_1052	12.2	4	Remote control design
AMI_20050204_1206	11.9	4	Remote control design
CMU_20050228_1615	12.0	4	Data collection for translation
CMU_20050301-1415	12.0	4	Transcription convention discussion
ICSI_20010531-1030	12.2	7	Meeting data collection
ICSI_20011113-1100	12.0	9	Staff meeting
NIST_20050412-1303	12.1	10	NIST ITL Diversity Committee meeting
NIST_20050427-0939	11.9	4	NIST computer support staff meeting
VT_20050304-1300	12.0	5	Tsunami relief planning
VT_20050318-1430	12.1	5	Scholarship selection committee
Total	120.4	56	
Unique speakers		46	

Lecture Room Meetings The *lectmtg* test set consisted of 150 minutes of lecture meeting excerpts from 16 different lectures recorded at Karlsruhe University[4]. The lectures were all technical language technology talks given by invited lecturers. Two types of excerpts were selected and transcribed by CMU: lecturer excerpts where the lecturer was the primary talker, and question/answer (Q&A) excerpts where the lecturer fielded questions from the audience. There were seventeen lecturer excerpts⁹ accounting for 89 minutes of data and twelve Q&A excerpts accounting for 61 minutes of data. Once the excerpts were selected, ELDA, KU and ITC-irst collaborated to annotate the data for the source localization task.

The audio sensors used in the *lectmtg* data were configured differently than the *confmtg* data. Only the lecturer and up to two audience members wore head-mounted, close-talking microphones. The rest of the audience was audible on the distant mi-

⁹ Two excerpts were selected from one of the meetings.

crophones. Four microphones were placed on the table in front of the lecturer and a fifth tabletop microphone was placed in the corner of the room. Four source localization arrays were mounted on each of the four walls of the room. Finally, a NIST Mark III array was placed directly in front of the lecturer.

2.5. Simultaneous Speech: STT Scoring

As previously noted, people often talk at the same time during meetings. The resulting overlapping speech represents a large challenge for speech technologies. Figure 1 is a cumulative histogram of the time within the test sets as a function of the number of active speakers¹⁰. It is evident from the graph that a large fraction of the time in each test set (~30% and 8% for the *confmtg* and *lectmtg* data sets respectively) involves simultaneous speech¹¹.

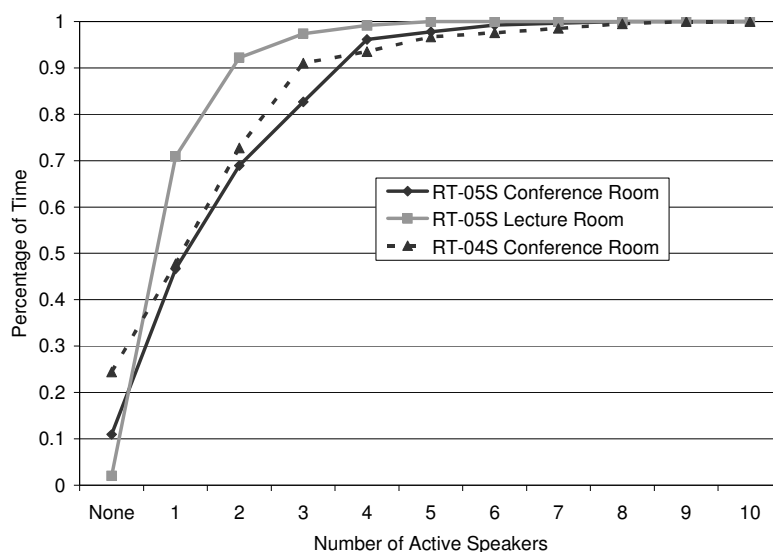


Figure 1 Cumulative histogram of testable time as a function of active speakers. The data points labeled 'None' indicate time in the test set where no one is talking

During the RT-04S evaluation, a prototype method was developed at NIST [2] to perform multi-stream STT scoring. For the RT-05S evaluation, a new alignment tool ASCLITE was developed in C++ and distributed to sites for use in the evaluation as

¹⁰ Active speakers calculated by segmenting the test set into independent regions for alignment where each region has no speakers crossing the region boundaries.

¹¹ Estimates published in 2004 for the RT-04S test set [2] were higher because silence regions in the test set were not included in the estimate.

part of the SCKT [6] toolkit. The techniques used by ASCLITE are described in [2]. The previous instantiation of SCLITE required a single reference and output word stream for alignment. ASCLITE represents an extension to support the alignment of multiple reference word streams (one for each speaker) to a single output word stream. Using this technique, system words are allowed to map to any reference word while maintaining the sequential ordering of words. This evaluation protocol enables the evaluation of single-stream STT systems using a multiple-speaker test set. Using ASCLITE, these systems can now be evaluated over segments of overlapping speech.

In order to limit the combinatorial explosion, the set of legal alignments is constrained to disallow the mapping of two reference words to each other. Even with this constraint, state-of-the-art computer system memory limits are exceeded. Experiments with the RT-04S and RT-05S test sets indicated that up to five simultaneous reference streams could be scored without exceeding memory limitations. Therefore, segments with greater than 5 simultaneous speakers were not evaluated. However, not much data was lost in this constraint. The majority of the test sets (98%, 97% and 100% of the RT-05S *confmtg*, RT-04S *confmtg*, and RT-05S *lectmtg* respectively) were able to be evaluated.

3. Results of the RT-05S Evaluation

3.1 RT-05S Evaluation Participants

The following table lists the RT-05S participants and the evaluation tasks each site took part in. In total there were nine sites submitting with three sites participating in two tasks.

Table 2 Summary of evaluation participants and the tasks for which systems were submitted

Site ID	Site Name	STT	SPKR	SAD	SLOC
AMI	Augmented Multiparty Interaction Program	X			
ICSI/SRI	International Computer Science Institute and SRI International	X	X		
ITC-irst	Center for Scientific and Technological Research				X
KU	Karlsruhe Univ.				X
ELISA Consortium	Laboratoire Informatique d'Avignon (LIA), Communication Langagière et Interaction Personne-Système (CLIPS), and LIUM		X	X	
MQU	Macquarie Univ.		X		
Purdue	Purdue Univ.			X	

TNO	The Netherlands Organisation for Applied Scientific Research	X	X
TUT	Tampere Univ. of Technology		X

3.2 Speech-To-Text (STT) Results

Two sites participated in the STT task, ICSI/SRI and AMI. They both ran their systems on both the confmtg and lectmtg data. While it is disappointing to see only two participants for this task, these two submissions represent many people's efforts across multiple research sites.

Appendix A contains the system performance graphs for the STT task. Figure 2 shows the WERs for the confmtg data set under the MDM audio input conditions as 46.9% and 38.4% for AMI and ICSI/SRI respectively. For ICSI/SRI, this represents a 33% relative reduction in WER from last year in which ICSI/SRI achieved a 53.4% WER (Figure 5). While the AMI WER was higher, this was the first public evaluation for the AMI system and a very good first showing. The lowest IHM WER was 25.9% compared to last year's 32.7% (Figure 5) which is a 20% relative reduction.

From Figure 2 The lectmtg data appears to be slightly harder than the confmtg data (13% and 40% relative for AMI and ICSI/SRI respectively) for the MDM microphone condition. However, the error rates comparing confmtg and lectmtg IHM systems are similar indicating that while the language difficulty may be equal, perhaps the challenges of distant microphones may not.

In Figure 4, the confmtg WERs by meeting shows no obvious performance outliers. However for the VT meetings, which are blind test data for the systems, ICSI/SRI did well on the VT meetings while AMI did not.

ICSI/SRI ran their system on the MSLA audio input condition on the lectmtg data. They achieved a 46.3% WER which is a 14% relative reduction from their MDM result. This is encouraging result. It is our opinion that as even more audio channels are used to record the speech, WERs will be reduced even further although the exact benefit can only be determined through experimentation.

When the lectmtg data is split into lecturer and Q&A subsets, there is no difference in performance for AMI. However ICSI/SRI did slightly better on the lecturer speech 51% as opposed to 58% on the Q&A speech.

3.3 Diarization "Who Spoke When" (SPKR) Results

Four sites participated in the SPKR task, the ELISA Consortium, ICSI/SRI, MQU and TNO. Appendix B contains the performance graphs for the SPKR task. The lowest DER for the primary systems on the confmtg data was 18.5% and 15.3% for the MDM and SDM audio input conditions respectively (Figure 6). Both scores were achieved by ICSI/SRI. These scores represent 20% and 32% relative reductions compared to RT-04S (Figure 10).

Oddly, this year's SDM error rates are lower than the MDM error rates. The difference is predominantly caused by the VT meetings where the SDM error rates are

64% and 24% lower than matched MDM DERs. For the rest of the meetings, the MDM and SDM scores are, by a sign test, not statistically different.

Figure 8 plots the MDM primary systems performance on the confmtg data set as a function of meeting id. The VT-20050318 meeting had the highest DER for all systems. Since the VT meetings are blind test data, this may be an indication of over-fitting to development data or, when coupled with the aforementioned low SDM error rates for the VT meetings, this may be an indication of differences in recording conditions. Neither can be proven with the existing data.

Like last year, the error rates for the full test set (the diamonds in Figure 7) are higher than just the non-overlapping speech.

The DERs for the lectmtg data set are in general lower than DERs for the confmtg data set. The best primary submission for all three audio input conditions is 12.2% for the ICSI/SRI team (Figure 6). This impressive result was achieved by declaring all testable speech as a single speaker without looking at the audio data. While the baseline is low, ICSI/SRI achieved a lower DER of 10.4% with a real system (Figure 7). In fact, most sites had contrastive systems with lower DERs so Figure 7 uses the lowest achieved DER for each data set/site/audio condition combination.

Not surprisingly, Figure 9 shows a marked difference in performance between the lecturer speech and the Q&A subset of the lectmtg data. System performance on the Q&A subset of the data is commensurate with the ELISA confmtg MDM results, but 27% higher for the ICSI/SRI system on the confmtg MDM data. It is difficult to say from the results whether or not the data was harder or if there was insufficient time for system development.

3.4 Diarization “Speech Activity Detection” (SAD) Results

Three sites participated in the SAD task: the ELISA consortium, Purdue and TNO. The lowest DER (from Figure 11) for the MDM condition on the confmtg data was 5.04% by TNO while ELISA’s error rate was 7.42%. These error rates are commensurate with the SPKR systems scored with the SAD evaluation criterion shown in Figure 7. ELISA achieved a 6.59% DER for the lectmtg data, MDM audio condition and the error rates for lecturer vs. Q&A subsets were roughly equal with 6.67% and 6.47% respectively

Purdue [15] was the only site to participate in the IHM audio input condition on the confmtg data. They achieved an SAD DER of 26.9%. As expected, SAD DERs are demonstrably higher for IHM than MDM or SDM conditions.

The SAD dry run was successful in that a system was built and tested with the evaluation infrastructure. However, the community is looking towards more challenging tasks, like the SPKR task, so SAD will not be an evaluation task in 2006.

3.5 Diarization “Source Localization” (SLOC) Results

Three sites participated in the SLOC task: ITC-irst, KU and TUT. The RMSE Error rates for the three systems were 309mm, 569mm, and 851mm respectively on the lecturer subset of the *lectmtg* test set.

The lowest error rate is impressive for the initial benchmark of this technology. The dry run was successful in that systems could be built to tackle the task and the evaluation methodology effectively measured system performance. However this level of performance is likely not sufficient for two reasons: (1) the systems were given lecturer-only speech, and (2) the required accuracy is likely to be lower than current performance.

These systems were given speech from a single talker, the lecturer. There was little competing speech and the lecturer's movements were typically constrained to the front of the room. As additional talkers are active and they move throughout the meeting space, error rates will degrade.

The community needs a good method to determine what constitutes sufficient accuracy; is it defined by geometrically measuring the person's location or by error rate reductions for consumers of SLOC system output? The current performance level is not sufficient for determining the exact person location. The average human male's neck to shoulder size is 235 mm¹² meaning the average error is beyond the person's body. The community is starting to research the uses of SLOC systems and the field has exciting possibilities.

4.0 Conclusions and Future Evaluations

In our opinion, the primary lesson learned from the RT-05S evaluation was that systems will do better as more sensors are collected; addition sensors will enable new tasks, like SLOC, and improve performance of existing tasks, like ICSI's WER reduction from additional distant microphones. The additional sensors also provide a rich opportunity for multimodal systems that blend the strengths of audio-based processing with video processing strengths. Indeed, this should be a new thrust in the meeting domain to support and experiment with sensor fusion techniques.

The successful evaluation of overlapping, simultaneous speech for the STT task is an indication that it is time to push for stream-based STT in the meeting domain. While it is invaluable to work on component technologies for a while, merging STT and SPKR is starting to make sense just like merging STT and segmentation in the early stages of the Broadcast News domain. With the advent of a plethora of audio channels, blind source separation (BSS) could easily support this task without modification to existing STT systems since BSS will deliver a single speaker's speech to the STT system.

The meeting recognition community is largely a volunteer group and therefore not encumbered with demands of program goals. The community should consider tackling unsolved problems in the STT field such as out-of-vocabulary (OOV) word detection. All of today's STT systems have finite vocabularies and including the capability of detecting OOVs would improve the usefulness of rich transcriptions to downstream users.

¹² The maximum shoulder width for a man's X-Large shirt according to the "Standards and Guidelines for Crochet and Knitting" by the Craft Yarn Council of America, <http://www.yarnstandards.com/s-and-g.pdf>

It is clear that the meeting recognition community is a vibrant and growing community. This year's goal should be to broaden the research base for technologies to pull in new participants with new ideas and energy.

5.0 Acknowledgements

NIST would like to thank everyone who donated meeting recordings for the evaluation, AMI, CMU, ICSI/SRI, VT and KU. Special thanks go to the sites that prepared the reference transcriptions and annotations: CMU, ELDA, KU, and LDC.

6.0 Disclaimer

These tests are designed for local implementation by each participant. The reported results are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U. S. Government.

References

1. Fiscus et. al., "Results of the Fall 2004 STT and MDE Evaluation", RT-04F Evaluation Workshop Proceedings, November 7-10, 2004.
2. Garofolo et. al., "The Rich Transcription 2004 Spring Meeting Recognition Evaluation", ICASSP 2004 Meeting Recognition Workshop, May 17, 2004
3. Spring 2005 (RT-05S) Rich Transcription Meeting Recognition Evaluation Plan, <http://www.nist.gov/speech/tests/rt/rt2005/spring/rt05s-meeting-eval-plan-V1.pdf>
4. Speaker Localization and Tracking – Evaluation Criteria, http://www.nist.gov/speech/tests/rt/rt2005/spring/sloc/CHIL-IRST_SpeakerLocEval-V5.0-2005-01-18.pdf
5. LDC Meeting Recording Transcription, <http://www ldc.upenn.edu/Projects/Transcription/NISTMeet>
6. SCTL toolkit, <http://www.nist.gov/speech/tools/index.htm>
7. "The ICSI Meeting Project: Resources and Research" A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters and B. Wrede, NIST ICASSP 2004 Meeting Recognition Workshop, Montreal
8. "The NIST Meeting Room Pilot Corpus", John S. Garofolo, Christophe D. Laprun, Martial Michel, Vincent M. Stanford, Elham Tabassi, LREC 2004
9. "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style", Susanne Burger, Victoria MacLaren, Hua Yu, ICSLP-2002
10. "Speech Activity Detection on Multichannels of Meeting Recordings", Zhongqiang Huang and Mary P. Harper, Proceedings from the RT-05 Workshop at MLML-05.

Appendix A: Speech-To-Text Results

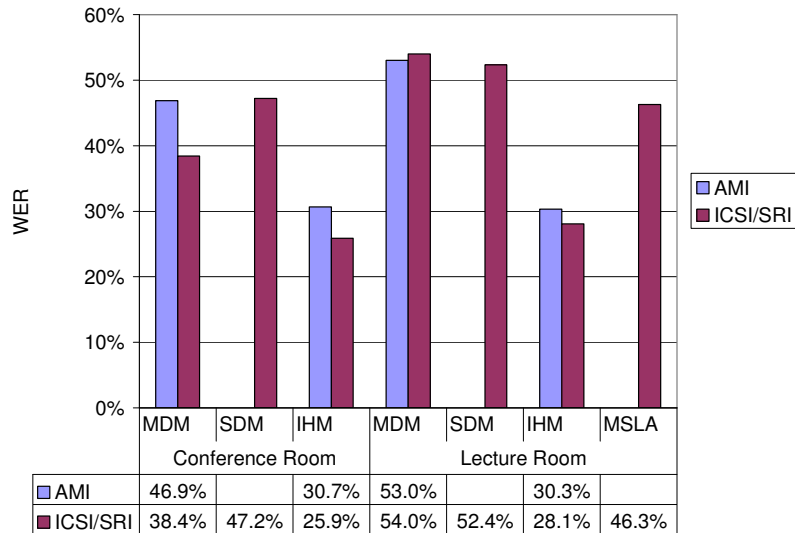


Figure 2 WERs for primary STT systems across test sets and audio input conditions

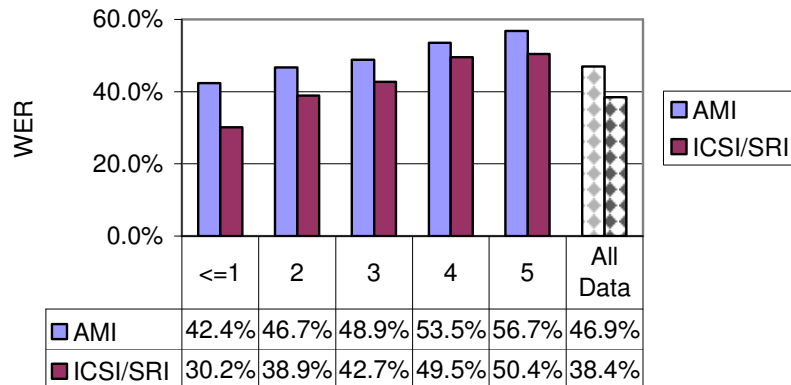


Figure 3 WERs for primary MDM STT systems as a function of the number of active speakers in a segment. The bars for ≤ 1 include regions where no one is talking and the final column "All Data" is the cumulative WER from Figure 2

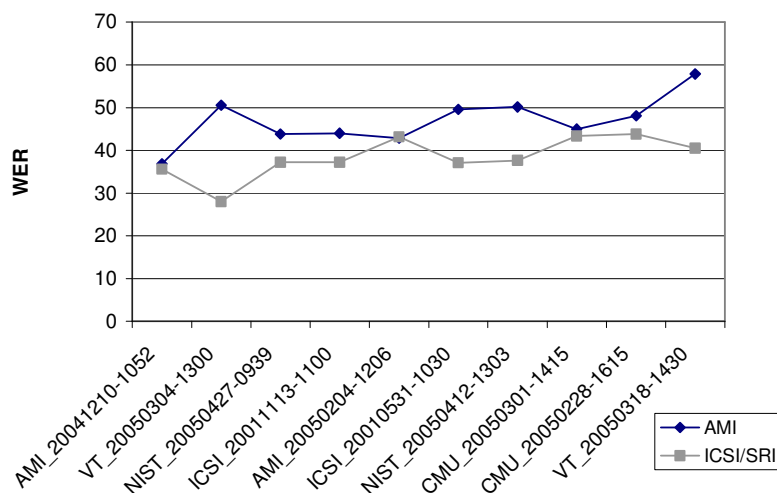


Figure 4 WERs for primary MDM STT systems broken down by meeting. Meetings are sorted by average WER

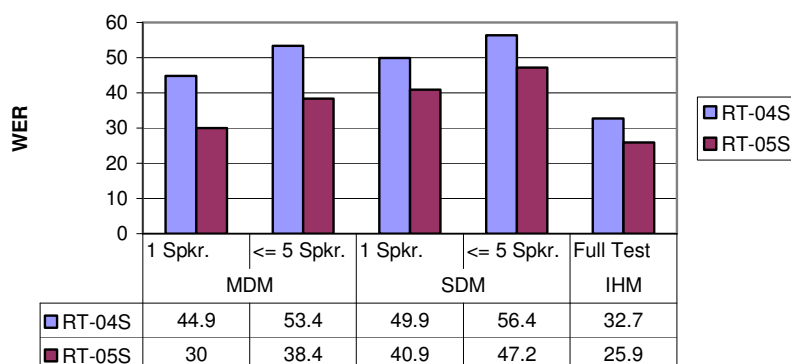


Figure 5 WERs for the best MDM and SDM STT systems from RT-04S and RT-05S. MDM and SDM results are broken down by “1 Spkr.” for non-overlapping speech, “<=5 Spkr.” which includes simultaneous speech, or ‘Full Test’ for the complete test set. The RT-04S systems were re-scored with ASCLITE

Appendix B: Diarization “Who Spoke When” (SPKR) Result

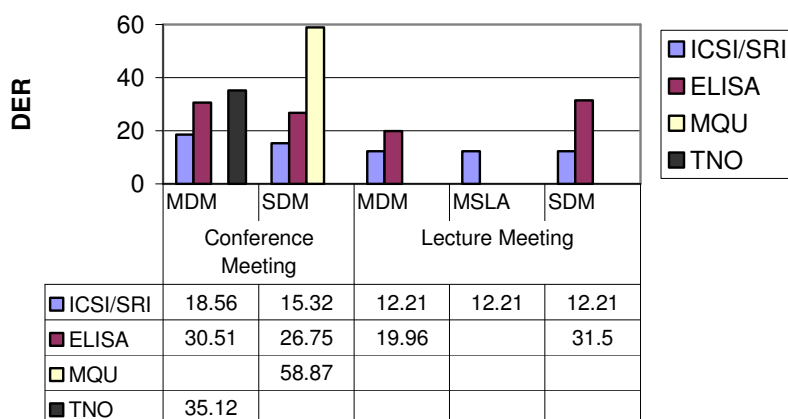


Figure 6 DERs for primary SPKR systems across test sets and audio input conditions

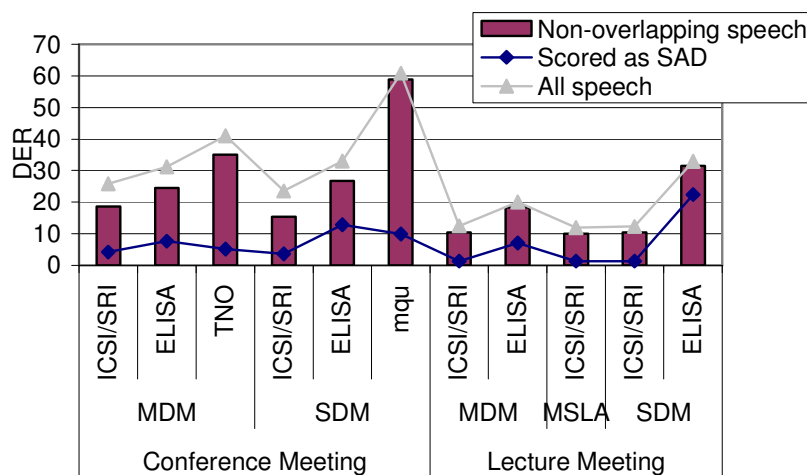


Figure 7 DERs for “Lowest Error Rate” systems from each site across test sets and audio input conditions. The *triangles* mark the DERs over all data including simultaneous speech. The *diamonds* mark the error rate of SPKR systems scored as SAD systems

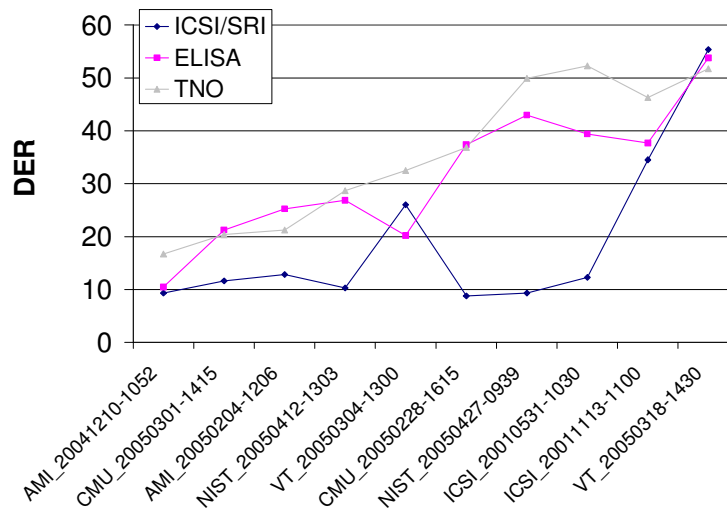


Figure 8 DERs for primary MDM SPKR systems broken down by meeting id. Meetings are sorted by average DER

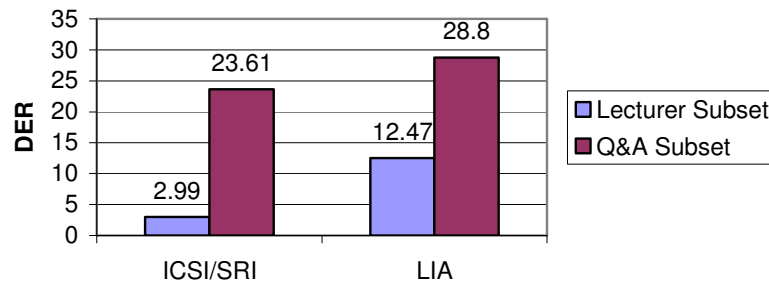


Figure 9 DERs for "Lowest Error Rate" systems from each site for the lectmtg data broken down by the *lecturer* and *Q&A* subsets

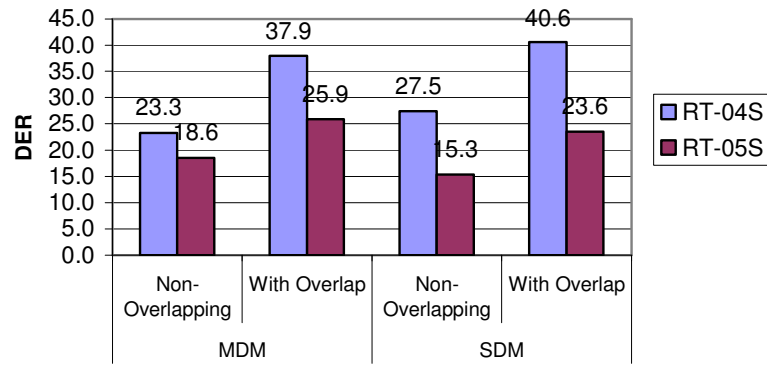


Figure 10 DERs for the best MDM and SDM SPKR systems from RT-04S and RT-05S

Appendix C: Diarization “Speech Activity Detection” Results

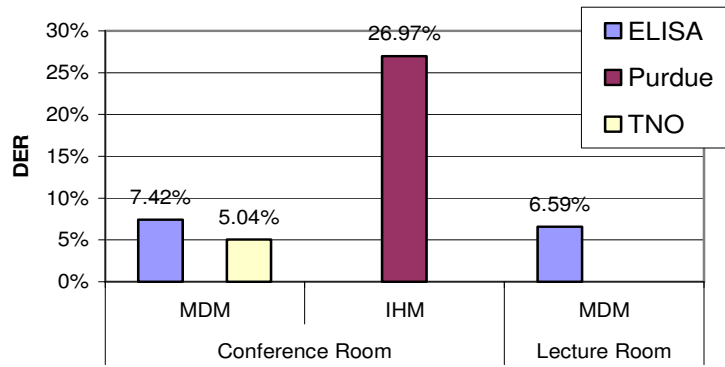


Figure 11 DERs for primary SAD systems across test sets and audio input conditions

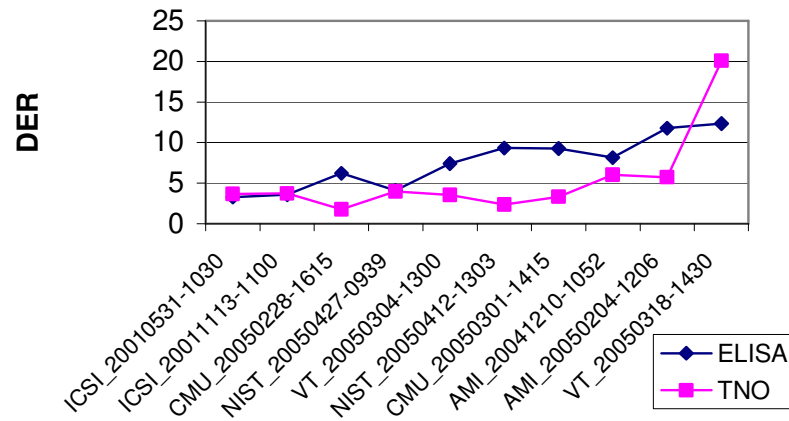


Figure 12 DERs for primary MDM SAD systems broken down by meeting id. Meetings are sorted by average DER